BBN TECHNOLOGIES
A Verizon Company

EARS The Blue Team
BBN LIMSI UPitt UW

BBN+UMD Rich Transcription System for Broadcast News

Daben Liu, Amit Srivastava, Francis Kubala,
Daniel Kiecza, Anson Ann, Jared Maguire,
Rich Schwartz
BBN Technologies

Matthew Snover, Bonnie Dorr
University of Maryland College Park

RT-03F Workshop

Washington, DC

13 November, 2003

Outline

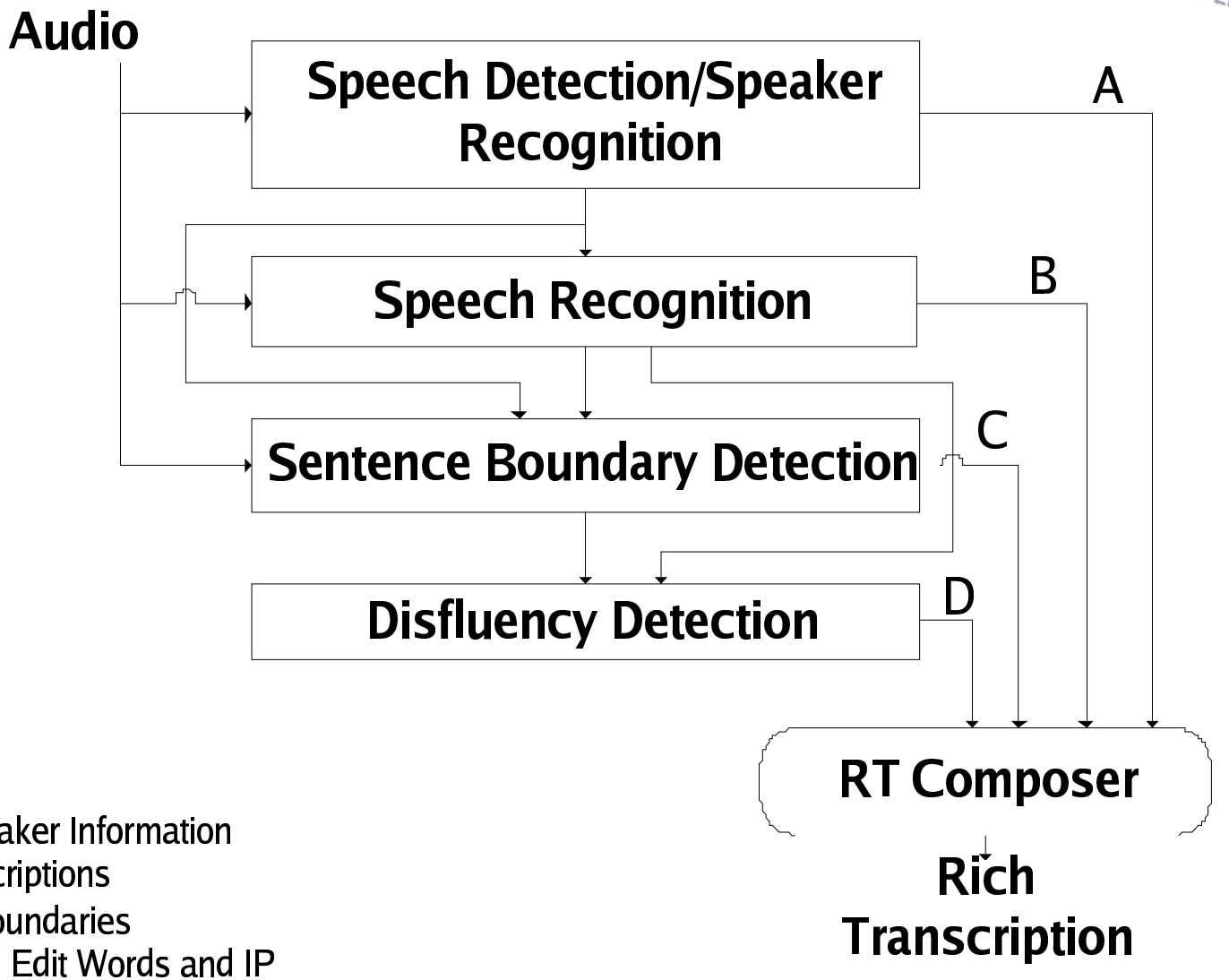
- **Broadcast News (BN) System**
  - Overview
  - Speaker Recognition
  - Sentence Boundary Detection

  (Speech-to-Text was from RT03S,
  Disfluency Detection will be given by Matt/Rich)

- **Evaluation Results**

- **Conclusion**

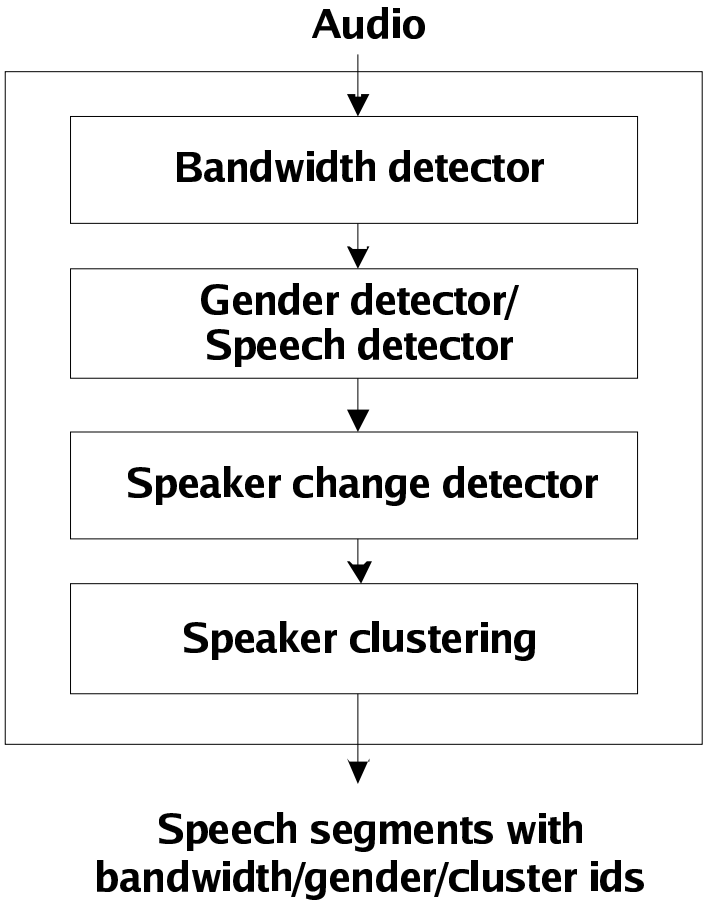
BN System Overview



A = Speech/Speaker Information
B = Word Transcriptions
C = Sentence Boundaries
D = Filler Words, Edit Words and IP

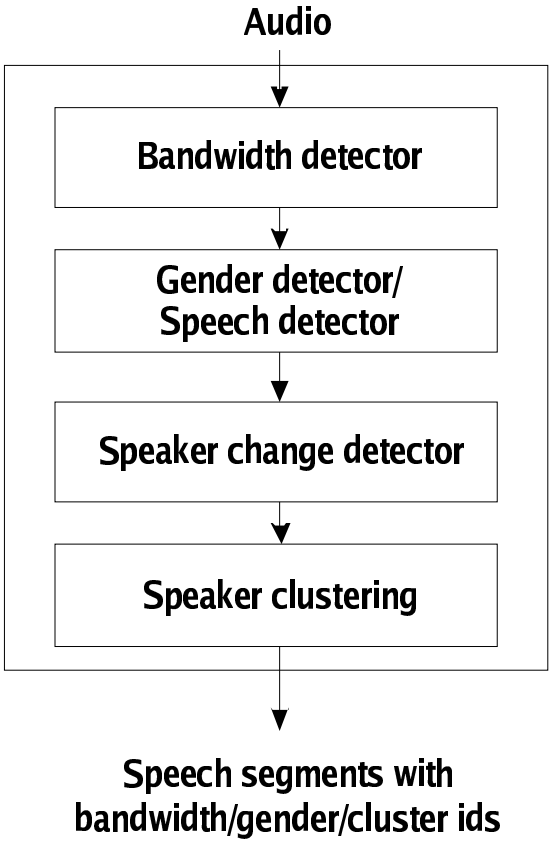
Speaker Recognition

- Improved version of the system used in December 2002 dry-run
- rteval_v2.3.pl as the scoring tool for development
- System Diagram for speech detection and speaker recognition:



Bandwidth Detection (new for RT03F)

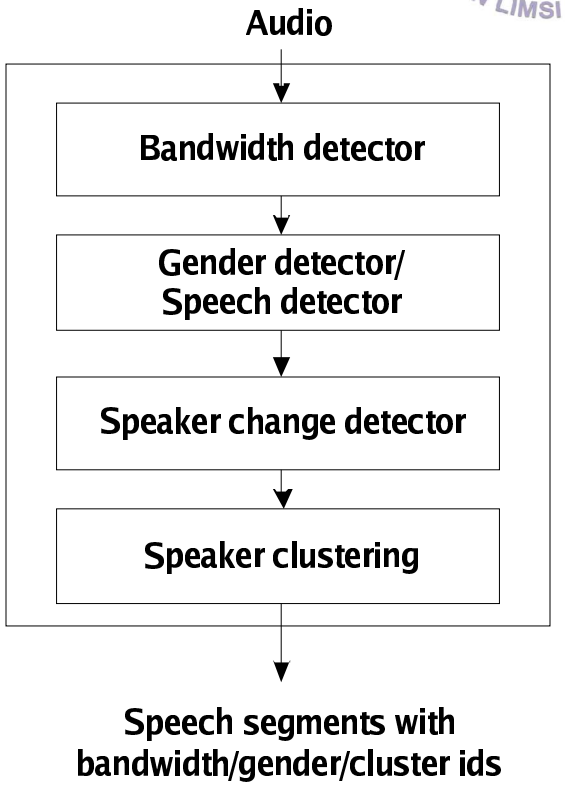
- 2-class GMM model for wide and narrow-band
  - Training Data from 3 languages, English, Chinese, Arabic
  - 20hrs for narrow-band, 40hrs for wide-band
  - 256 GMM components
  - 20-state HMM: 0.2sec minimum duration
  - Viterbi decode
- Benefit
  - Simpler model
  - More general and robust
  - 0.2% improvement on speaker recognition (SR) score

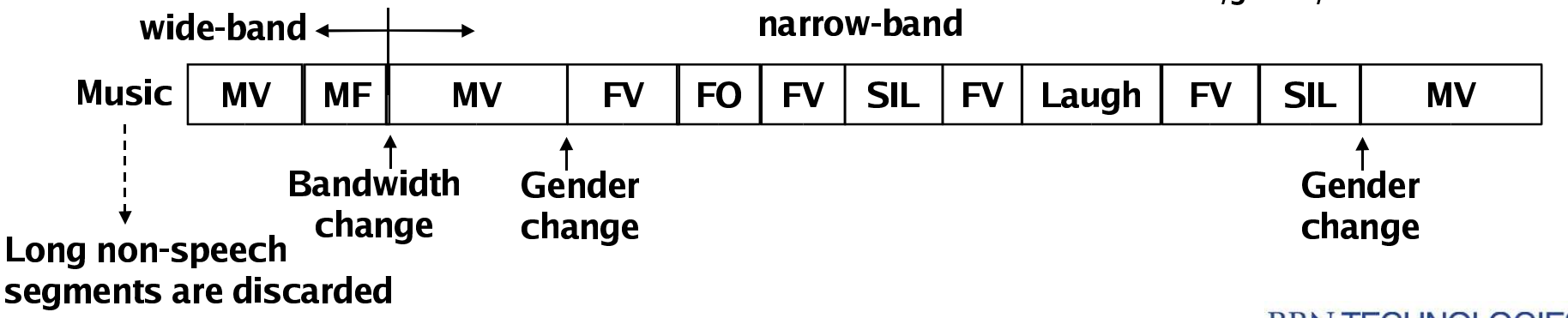


Gender/Speech Detection (unchanged for RT03F)

- Detect within bandwidth-specific segments
- Phoneme decode with 11 classes
  - speech phones: MV, MF, MO, FV, FF, FO
  - non-speech phones: music, silence, breath, lip-smack, laughter
- Training from Hub4 98, 80hrs, male:female =~ 2:1
- Output:

  sequence of broad phoneme classes



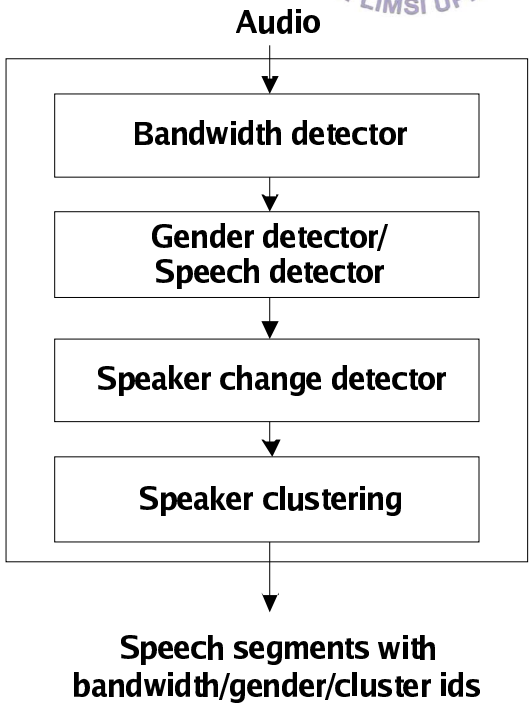


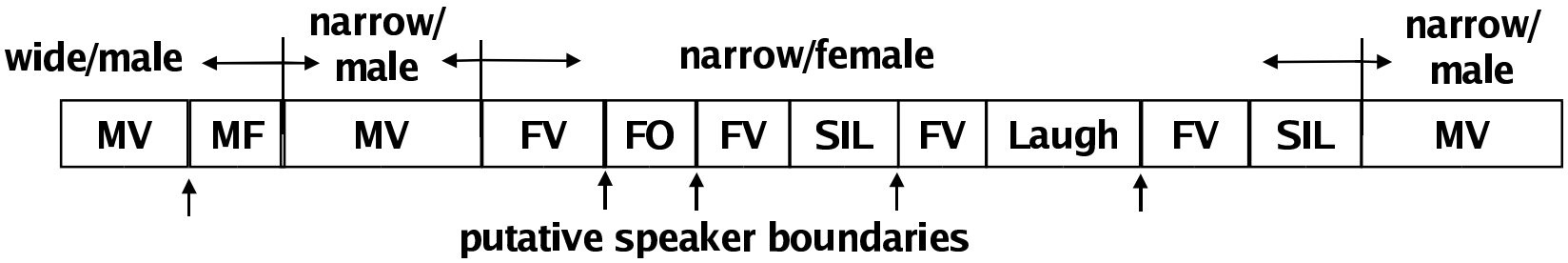
Speaker Change Detection (unchanged for RT03F)

- Goal was to find speaker changes within bandwidth-gender specific speech segments
- Hypothesize speaker change on every phoneme class boundary. On average, reduce computation by a factor of 10
- Generalized Likelihood Ratio (GLR) test with duration penalty:

$$= \frac{L(z;\mu_z,Z_z)}{L(x;\mu_x,Z_x)L(y;\mu_y,Z_y)} \cdot \left(\frac{1}{N}\right)$$

- Non-speech frames not used for GLR tests
- Biased to find more changes on non-speech phonemes

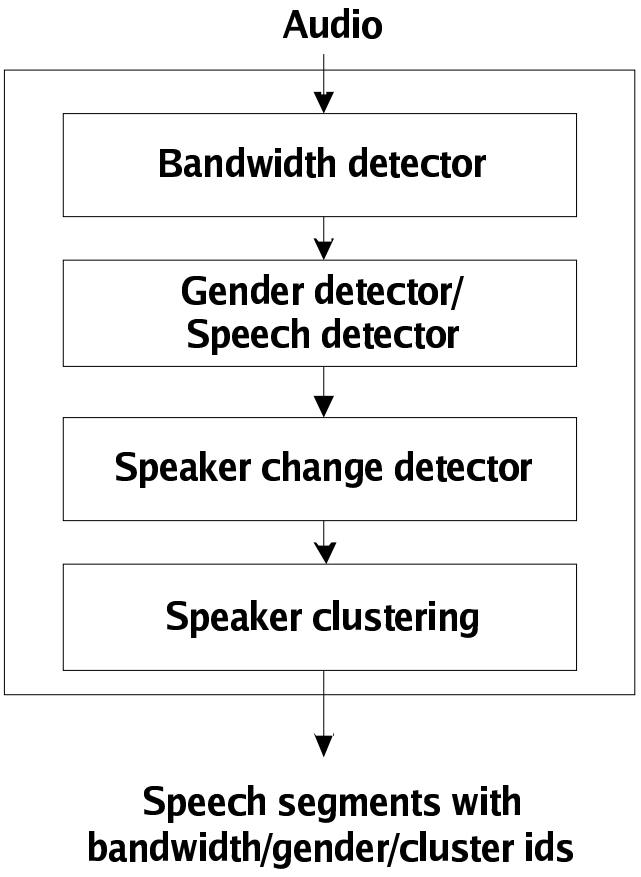




D. Liu, F. Kubala, "Fast Speaker Change Detection for Broadcast News Transcription and Indexing," *EUROSPEECH'99*, Budapest, Hungary, Volume 3, Page 1031-1034, September 5-9, 1999

Speaker Clustering (improved for RT03F)

- Online speaker clustering
  - Clustering decisions are made on the fly
  - Causal process with no latency. Decision cannot be changed later
  - Cannot change bandwidth and gender boundaries
  - Can change speaker boundaries detected by speaker change detection
- Benefit compared to offline hierarchical-style speaker clustering
  - Simpler approach
  - Consistently more accurate
  - Run faster
  - No stopping criterion is needed
- Recent improvement
  - Distance measure uses the same duration-penalized GLR as used in speaker change detection. (Previous system omitted the penalty term)
  - First and Second order cepstral derivatives added as new features



D. Liu, F. Kubala, "Online Speaker Clustering," *ICASSP'03*, Hong Kong, May, 2003

BBN TECHNOLOGIES
A Verizon Company

Improvement Summary for Speaker Recognition

- STT segmentation from RT03S evaluation as the baseline
- Dev03F: 1.5hrs from ABC, NBC, CNN
- Scored by rteval_v2.3.pl

Improvements	SR	RT 1	RT03
1. STT segmentation (baseline)	32.2	12.7	42.7
2. Bandwidth viterbi decode	32.0	12.7	42.5
3. Online speaker clustering*	30.9	12.7	40.5
4. Turning clustering parameters	27.9	12.7	39.2
5. Duration-penalized GLR for clustering	26.6	12.7	38.0
6. Improved STT from RT03S	25.1	11.2	35.0
7. Add derivatives as clustering feature	23.4	11.2	32.4
Relative improvement to baseline	25%	12%	24%

*The initial parameters for 3. was tuned on Hub4 1996 evaluation data, with reference segmentation

Conclusions

- Tuning resulted in the biggest gain of 3% absolute
- Online speaker clustering was 1.1% better than offline speaker clustering in terms of SR scores
- Cepstral derivatives gave a big gain of 1.7% absolute
- RT03 improvement tracks the SR improvement

Sentence Boundary Detection (SBD)

- System is the same as that used in CTS, except for the following differences
  - Sentence boundary decisions were made on bandwidth, gender, and speaker boundaries detected by speaker recognition
  - Linguistic subsystem was only word-based. Part-of-Speech (POS) was not implemented for BN
  - No system combination

Training Data

- **Acoustic training**
  - 17 hours of MDE training data released by LDC, which conforms to "MDE Annotation Spec v5"
  - 70 hours of Hub4 acoustic training data
- **Language model training**
  - All acoustic training data
  - TDT4 transcripts
  - 3 million words
  - Additional PSM data did not help

Improvement Summary for SBD

Baseline uses a silence chopper on the gender decode output

chop on longer silence first
average sentence duration was 4 second

Improvement	SBD	RT1	RT03
Silence chopper (baseline)	66.5	11.2	32.4
SBD with CTS settings	64.0	11.2	32.1
SBD parameters tuning for BN	61.8	11.2	32.1
Cleanup language model training	61.0	11.2	31.8
Add PSM language model training	62.3	11.2	32.1
More Neural Net training epochs (from 130 to 143)	58.5	11.2	31.6
Relatively improvement	12%	-	2%

Conclusions

- **Parameter tuning for BN resulted in a gain of 2.2%**
- **More epochs gave a big gain of 2.5%. However no significant gain was observed beyond 143rd epoch**
- **RT03 score was not sensitive to SBD score changes due to the fact that SBD had a much smaller denominator**

Evaluation Results

- **Eval03F: 1.5hrs from PRI, VOA, MSNBC**

Test	SBD	Edit	Filler	IP	SR	RT1	RT03
Dev03F	58.5	98.7	81.4	96.2	23.4	11.2	31.6
Eval03	63.8	94.5	78.8	85	15	11.7	24.3

Conclusions

- **Good**
  - **Very good speaker recognition result. Apparently Eval03F set was easier than Dev03F set**
  - **Edit performance was better for Eval03F**
  - **RT03 was also much better, mainly due to better SR**
- **Could be better**
  - **SBD was about 10% relatively worse for Eval03F**
  - **Filler performance was about the same, but … (next slide)**

Problem with filled pauses

- Filled pause detection solely depends on STT, which is not tuned to recognize filled pauses. For STT, pauses are optionally deletable
- Dev03F and Eval03F are very different on filled pauses (why?)

set	#ref	#hyp	#corr	#ins
Dev03F	45	107	36 (80%)	71 (158%)
Eval03F	204	280	176 (86%)	104 (51%)

- Decision made based on Dev03F
  - Stripped out all the uhs before submission (~90% of filled pauses hypothesized). Filler errors dropped from 129% to 81%
  - For Eval03F, the effect is the opposite. If uhs are preserved, filler error would be 57%, rather than 79%. Most of other conditions also gained

Test	SBD	Edit	Filler	IP	SR	RT1	RT03
Dev03F	58.5	98.7	**81.4**	96.2	23.4	11.2	31.6
Eval03F	63.8	94.5	**78.8**	85.6	15.1	11.7	24.3
uh-preserved Dev03F	59.4	98.7	**128.6**	125.5	23.2	11.1	31.6
uh-preserved Eval03F	64.0	93.9	**57.2**	70.1	14.1	10.9	23.6

Conclusion

- We participated in RT03F evaluation for all conditions in BN
- The final RT-03 error is 24.3%. STT RT1 error (11.7%) accounts for less than 50% of the total error. Most of the errors would be due to SR errors.
- Most CTS technologies applicable to BN
- We had less than 1 person-month effort on BN system development for this evaluation. Most of time spent on understanding the new scoring tools and new training data
- We hope the Dev data could be statistically close to Eval data, especially on those features to be evaluated.